PROCEEDINGS OF SPIE

SPIEDigitalLibrary.org/conference-proceedings-of-spie

Towards a quantitative analysis of class activation mapping for deep learning-based computer-aided diagnosis

Kang, Hanul, Park, Ho-min, Ahn, Yuju, Van Messem, Arnout, De Neve, Wesley

Hanul Kang, Ho-min Park, Yuju Ahn, Arnout Van Messem, Wesley De Neve, "Towards a quantitative analysis of class activation mapping for deep learning-based computer-aided diagnosis," Proc. SPIE 11599, Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment, 115990M (15 February 2021); doi: 10.1117/12.2580819

SPIE.

Event: SPIE Medical Imaging, 2021, Online Only

Towards a Quantitative Analysis of Class Activation Mapping for Deep Learning-based Computer-Aided Diagnosis

Hanul Kang[a], Ho-min Park[a,b], Yuju Ahn[a], Arnout Van Messem[a,c], and Wesley De Neve[a,b]

[a]Center for Biotech Data Science, Department of Environmental Technology, Food Technology and Molecular Biotechnology, Ghent University Global Campus, Incheon, Korea
[b]IDLab, Department of Electronics and Information Systems, Ghent University, Ghent, Belgium
[c]Department of Mathematics, University of Liège, Liège, Belgium

ABSTRACT

Class Activation Mapping (CAM) can be used to obtain a visual understanding of the predictions made by Convolutional Neural Networks (CNNs), facilitating qualitative insight into these neural networks when they are, for instance, used for the purpose of medical image analysis. In this paper, we investigate to what extent CAM also enables a quantitative understanding of CNN-based classification models through the creation of segmentation masks out of class activation maps, hereby targeting the use case of brain tumor classification. To that end, when a class activation map has been created for a correctly classified brain tumor, we additionally perform tumor segmentation by binarization of the aforementioned map, leveraging different methods for thresholding. In a next step, we compare this CAM-based segmentation mask to the segmentation ground truth, measuring similarity through the use of Intersection over Union (IoU). Our experimental results show that, although our CNN-based classification models have a similarly high accuracy between 86.0% and 90.8%, their generated masks are different. For example, our Modified VGG-16 model scores an mIoU of 12.2%, whereas AlexNet scores an mIoU of 2.1%. When comparing with the mIoU obtained by our U-Net-based models, which is between 66.6% and 67.3%, and where U-Net is a dedicated pixel-wise segmentation model, our experimental results point to a significant difference in terms of segmentation effectiveness. As such, the use of CAM for the purpose of proxy segmentation or as a ground truth segmentation mask generator comes with several limitations.

1. INTRODUCTION

Deep learning using Convolutional Neural Networks (CNNs)[1] is highly successful in classifying and segmenting images. As a result, the domain of healthcare has a strong interest in the use of CNNs for diagnosing diseases,[2–7] with the correct treatment of a disease starting with an accurate diagnosis. Often, such a diagnosis heavily relies on medical imaging techniques, including optical camera photographs, X-rays, Magnetic Resonance Imaging (MRI) scans, and Positron Emission Tomography (PET) scans. With the help of CNNs, an ever-increasing number of medical images can be analyzed, aiding medical experts in decision-making by significantly reducing the time and effort needed for the detection and classification of abnormalities. However, CNNs are so-called black-box predictive models, providing a limited understanding of their internal working and the rationale behind the predictions made. Fortunately, the introduction of Class Activation Mapping (CAM)[8] made it possible to gain a better insight into the way CNNs process images, and ever since, numerous studies have leveraged CAM to better understand how well CNNs are doing in analyzing medical images.[2,3,5,6] However, most of these studies only use CAM for qualitative model assessment, through the creation of heatmaps. In this paper, we outline an approach that uses CAM for quantitative model analysis, measuring how well CAM can be additionally used to segment a brain tumor in MRI images, given that the type of brain tumor has been correctly determined by the underlying CNN-based classification model.

The first two authors contributed equally. Send correspondence to Ho-min Park: homin.park@ugent.be

Medical Imaging 2021: Image Perception, Observer Performance, and Technology Assessment, edited by Frank W. Samuelson, Sian Taylor-Phillips, Proc. of SPIE Vol. 11599, 115990M
 · doi: 10.1117/12.2580819

2. BACKGROUND

2.1 Medical imaging

Medical imaging aims at visualizing interior body parts to aid in clinical analysis and diagnosis of diseases by medical experts. Over the past few years, the number of available medical images has increased substantially, thanks to the development of novel imaging techniques and an expansion in digital storage.[9] Furthermore, recent advancements in computer vision, and in the area of deep learning in particular, have enabled automatic analyses of vast amounts of health data. In the field of medical imaging, deep learning is typically used for the detection, classification, and segmentation of abnormalities. Moreover, generating visualizations in support of the predictions made is an important functionality, making it possible for medical experts to better understand why certain decisions have been made. In this paper, we investigate to what extent CAM, a qualitative visualization technique, can also be used for quantitative model analysis.

2.2 Class Activation Mapping

CAM is a visualization technique that highlights discriminative image regions that are relevant to a particular class, and where these image regions are used by a model to identify the given class.[8] To generate a class activation map, a so-called Global Average Pooling (GAP) layer needs to be present after the final convolutional layer of a particular model. Such a GAP layer produces a single number by taking the sum of all values in a feature map. The architectural requirement of having a GAP layer after the final convolutional layer allows a neural network to consider the average activation of the entire image, rather than the activation of a specific image region, as is for instance the case when making use of a Global Max Pooling (GMP) layer. The weights used to generate a classification outcome are projected back onto the final feature maps and the weighted sum of these feature maps subsequently produces a class activation map that takes the form of a heatmap. The weights used specifically indicate the importance of each image region in determining the class of interest. In a next step, the generated heatmap is up-sampled to the size of the given input image, and by then overlaying this heatmap on the input image at hand, the location of the most informative image regions can be found. In what follows, we provide a more formal description of the aforementioned steps.

Assume that the kth feature map of the output of a CNN (used as a feature extractor) is denoted as $A^{(k)} \in \mathbb{R}^{u \times v}$. For $i = 1, \ldots, u$ and $j = 1, \ldots, v$, let $A_{ij}^{(k)}$ denote the element of the matrix $A^{(k)}$ corresponding to the ith row and the jth column. That way, the classification score of the cth class is obtained as follows:

$$\hat{y}^{(c)} = \sum_{k} w_k^{(c)} \sum_{i,j} A_{ij}^{(k)}, \tag{1}$$

where $w_k^{(c)} \in \mathbb{R}$ is the weight of the edge connecting the kth feature map and the cth class, and $\sum_{i,j} A_{ij}^{(k)}$ is the sum over all elements in the kth feature map. The elements of the class activation map $S^{(c)} \in \mathbb{R}^{u \times v}$ that corresponds to the cth class can then be obtained as follows:

$$S_{ij}^{(c)} = \sum_{k} w_k^{(c)} A_{ij}^{(k)}. \tag{2}$$

Since $S^{(c)}$ has gone through several pooling layers in the CNN used, this class activation map comes with a smaller size than the original input image $I \in \mathbb{R}^{w \times h}$. We denote by $H^{(c)} \in \mathbb{R}^{w \times h}$ the matrix obtained after up-sampling the matrix $S^{(c)}$.

As previously mentioned in Section 1, CAM is a technique that facilitates a visual understanding of what a predictive model looks at in input images when solving a task like computer-aided detection and classification of abnormalities in medical images, making it possible to assist medical experts in diagnosing diseases.[10] In this paper, we additionally leverage CAM for the purpose of quantitative model analysis, working with CNN-based classification models that take as input brain MRI images.

Table 1. CNN-based models for computer-aided diagnosis in medical images and corresponding visualization methods.

Authors	Use case	Visualization
Rajpurkar *et al.*[2]	Pneumonia detection using chest X-rays	CAM
Bien *et al.*[3]	Knee injury classification using MRI scans	CAM
Han *et al.*[4]	Skin disorder classification using clinical skin images	Grad-CAM
Nguyen *et al.*[5]	Eye tumor segmentation using MRI scans	CAM
Kiani *et al.*[6]	Liver cancer detection using biological tissue images	CAM
Kim *et al.*[7]	Glaucoma classification using fundus images	Grad-CAM

2.3 Class Activation Mapping in medical imaging

As shown in Table 1, CAM and a generalized version called Gradient-weighted CAM (Grad-CAM)[11] are regularly used as visualization techniques for CNN-based models.

Rajpurkar *et al.*[2] developed CheXNet, a 121-layered CNN, for diagnosing pneumonia in chest X-ray images. CheXNet offers superior levels of effectiveness, compared to conventional methods relied on by experienced radiologists. Indeed, CheXNet can correctly detect pathology, also producing the probability of having pneumonia. Furthermore, CAM is used as a qualitative measure to localize the most probable area of pathology in a given input image.

On a similar note, Bien *et al.*[3] proposed MRNet, a deep learning approach to reduce diagnostic errors and to speed up time-consuming analyses of knee MRI images. Indeed, automatic analyses by deep learning methods can aid in diagnosis and help to prioritise high-risk patients, with these methods learning features from input images provided during training. Moreover, the study by Bien *et al.* also leveraged CAM as a visualization tool, making it possible to observe which parts of an input image have the most significant influence on the predictions produced.

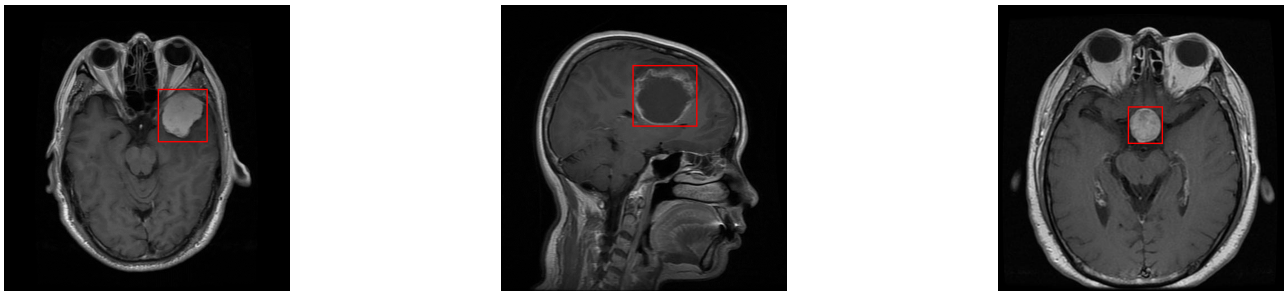
Han *et al.*[4] developed ResNet-152, a deep learning tool for initial screening of skin cancer, making it possible to classify twelve different skin-related diseases. Grad-CAM is used as a tool to further understand the model predictions and to create visual explanations, by taking advantage of gradient information flowing into the last convolutional layer.

For accurate diagnosis and treatment planning of eye tumors, Nguyen *et al.*[5] developed a segmentation model that combines two CNNs: ResNet and U-Net. The developed segmentation model is used to perform quantitative analyses of eye tumor tissues, supporting health workers in planning efficient therapies for patients. CAM is used for the localization and segmentation of malicious tissue in input images. Specifically, class activation maps, as generated for CNNs that aim at detecting eye tumors, are further refined, using these maps for the purpose of training U-Net-based segmentation models.

Kiani *et al.*[6] introduced a deep learning-based assistant to aid pathologists with the classification of liver cancer, creating an easily accessible diagnostic tool. This study also includes an evaluation of the influence of model effectiveness on diagnoses made by pathologists. Furthermore, CAM is used as an explanatory tool for the generated predictions.

To overcome the labour-intensive nature of image processing and manual screening, Kim *et al.*[7] proposed a model for computer-aided glaucoma detection. In particular, the proposed model is able to assist with the diagnosis and localization of glaucoma in eye fundus images. Furthermore, Grad-CAM is used to localize glaucomatous regions in an input image, without having to provide explicit segmentation information to the model trained.

In summary, all of the above studies leverage CAM or Grad-CAM as tools to visually localize malignancies and abnormalities in the images used. Moreover, in most studies, CAM and Grad-CAM are only used for qualitative model analysis, helping experts in achieving a better understanding of the predictions made. In the remainder of this paper, we set out to investigate the applicability of CAM as a quantitative tool.



(a) Meningioma (b) Glioma (c) Pituitary tumor

Figure 1. Example image for each type of brain tumor. A red box is used to delineate a tumor.

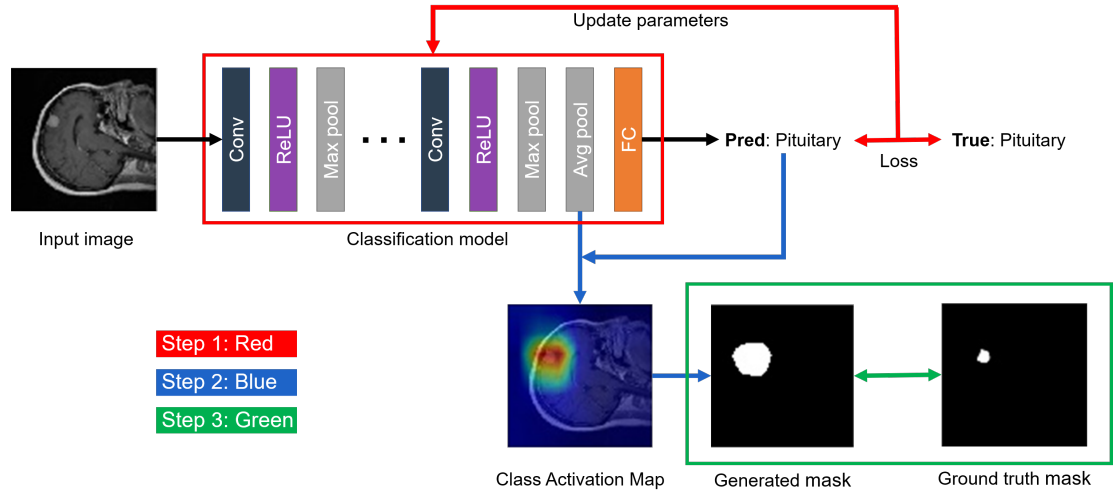


Figure 2. Overview of the proposed approach, leveraging a CNN that consists of convolutional layers (Conv), rectified linear unit layers (ReLU), pooling layers (average and maximum), and a fully connected layer.

3. METHODS

3.1 Dataset description

We make use of a brain tumor dataset[12] that was acquired between 2005 and 2010 at two Chinese hospitals, namely the Nanfang Hospital in Guangzhou and the General Hospital of Tianjing Medical University. This dataset consists of 3,064 T1-weighted brain MRI images stemming from 233 patients. Each image has been given a multi-class brain tumor label. Moreover, each image is accompanied by its corresponding segmentation mask (ground truth). A distinction is made between three types of tumors: (1) meningioma (708 images), (2) glioma (1,426 images), and (3) pituitary tumor (930 images).

An example of each tumor type can be found in Figure 1. A meningioma, which can be found in-between the skull and the brain, originates from the arachnoid cells that surround the brain. A glioma, which can be found anywhere inside the skull, originates from the glial cells that deliver the substances needed for neuron functioning. A pituitary tumor is usually found near the center of the brain, occurring in the pituitary gland that governs the secretion of hormones. Furthermore, MRI images are, in general, divided over three anatomical planes: (1) the sagittal plane, (2) the axial plane, and (3) the coronal plane. However, the adopted dataset does not contain any information about the anatomical plane used. Note that the images presented in Figure 1(a) and Figure 1(c) were created along the axial plane, while the image presented in Figure 1(b) was created along the sagittal plane.

3.2 Deep learning models

3.2.1 Classification models

The first step, training a CNN-based classification model, is indicated in red in Figure 2. When an image is given as an input to a classification model, the model subsequently predicts the type of tumor (label) present in the image. Next, the difference between the predicted label and the true label (loss) is calculated. This loss is then used for updating the parameters in the model, making it possible to obtain more accurate predictions, and thus a lower loss. We use six classification models: (1) AlexNet,[1] (2) VGG-16,[13] (3) VGG-19,[13] (4) Modified AlexNet, (5) Modified VGG-16, and (6) Modified VGG-19. A GAP layer is added to AlexNet before the fully connected layer. For each of the VGG models, the target output size of the image after GAP is changed from 7×7 to 1×1 in order to be able to create a class activation map in the next step. The Modified AlexNet and VGG models are implementations of AlexNet*-GAP[8] and VGGnet-GAP,[8] respectively. The final max pooling layer before the average pooling layer is removed and replaced with two additional convolutional layers for Modified AlexNet, and with one additional convolutional layer for the Modified VGG models. Such a replacement aims at improving the localization ability of the models used. To mitigate the risk of overfitting, we applied transfer learning,[14] initializing the model parameters with values that have already been learned from ImageNet.[15]

The brain tumor dataset used is imbalanced in nature. To overcome imbalance, undersampling of the majority class or oversampling of the minority class can be used. However, we did not use undersampling due to an insufficient amount of data and we did not make use of oversampling due to the possibility of overfitting. Instead, we made use of 5-fold cross-validation so that each image is used exactly once in a validation set. In particular, since the total number of images (3,064) is not divisible by five, our dataset was split into four subsets of 613 images, and one subset of 612 images (the proportions of the different classes were kept the same in each subset). Each subset was then used four times as a training set, for the purpose of parameter tuning, and once as a validation set, for the purpose of model evaluation. As a result, a total of five (training, validation)-combinations were obtained for each model, using the average effectiveness (as measured through accuracy, precision, recall, and F1 score; see Section 3.4.1) to quantify the final model performance. That way, a total of six classification models were trained.

3.2.2 Semantic segmentation models

Next to classification models, we trained models for semantic segmentation, taking advantage of U-Net,[16] a neural network often used in the area of biomedical image segmentation. We made use of VGG-16 and VGG-19 as encoders for U-Net[16] and applied transfer learning.[14] We relies on Dice loss, which measures the amount of overlap between a predicted mask and a ground truth mask, for parameter tuning. Since Dice loss is known to be insensitive to label imbalance, it is widely used for constructing segmentation models.[17] Finally, by comparing the segmentation effectiveness of our U-Net-based models with the segmentation effectiveness of our classification models, as implemented by binarization in a later step in Section 3.3, we are also able to evaluate the segmentation effectiveness (using IoU, as discussed in Section 3.4.2) of these classification models.

3.3 Generating proxy segmentation masks using CAM

The second step, indicated in blue in Figure 2, consists of generating the class activation map and the corresponding segmentation mask for a particular input image. The class activation map is created by combining the output of the average pooling layer and the predicted tumor type.

Furthermore, to obtain a binary segmentation mask out of a class activation map, we made use of thresholding. Specifically, global thresholding, which is also known as simple thresholding, is a method that can be used to extract an object of interest (i.e., a brain tumor) from the background by making use of a single fixed threshold value T for all pixels in an image of size $w \times h$. As such, the elements of the resulting binary segmentation mask $M \in \mathbb{R}^{w \times h}$ can be defined as follows:

$$M_{ij} = \begin{cases} 255, & \text{if } H_{ij}^{(c)} > T \\ 0, & \text{if } H_{ij}^{(c)} \leq T \end{cases}, \quad i = 1, \ldots, w;\ j = 1, \ldots, h. \tag{3}$$

In other words, if the pixel value in a class activation map is greater than the threshold value T, it is set to white (255), otherwise it is set to black (0). The chosen threshold values T are as follows: 200, 225, 90th percentile (Q90), and 95th percentile (Q95). Another method that we applied is Otsu thresholding,[18] which automatically chooses a threshold value for each pixel based on the neighboring region within an image. While thresholding based on 200 and 225 simply uses a manually chosen value for T, thresholding based on Q90, Q95, and Otsu determines the value automatically, taking into account the pixel distribution of an image.

3.4 Evaluation metrics

3.4.1 Classification

The classification effectiveness of the models is evaluated using the following metrics: accuracy, precision, recall (sensitivity), and F1 score. For the accuracy, we used the proportion of correctly classified cases, i.e., those cases for which the prediction $\hat{y}_i$ corresponds to the true value y_i. For the other metrics, we made use of the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) that can be found in the confusion matrix created for our three-class classification problem. Since the image distribution over the different classes is only slightly skewed (meningioma: 23.1%, glioma: 46.5%, and pituitary tumor: 30.4%), we opted to make use of the macro-average to determine the final precision, recall (sensitivity), and F1 score values. In other words, the aforementioned metrics were calculated for each class separately, after which the average over the three classes was taken. All four metrics were calculated relying on the implementation that is available in scikit-learn.*

$$Accuracy = \frac{1}{n_{samples}} \sum_{i=1}^{n_{samples}} \mathbb{1}(\hat{y}_i = y_i) \tag{4}$$

$$Precision = \frac{TP}{TP \ + \ FP}, \tag{5}$$

$$Recall = \frac{TP}{TP \ + \ FN}, \tag{6}$$

$$F1 \ score = 2 * \frac{Recall \ * \ Precision}{Recall \ + \ Precision}. \tag{7}$$

3.4.2 Segmentation

The segmentation effectiveness of the different models is evaluated using the Intersection over Union (IoU). IoU, also known as the Jaccard Index, represents the area of intersection over the area of union between the segmentation mask predicted by CAM (M) and the corresponding ground truth mask (Ω). The average of all IoU values obtained over all classes is called the mean IoU (mIoU).

$$IoU = \frac{|M \ \cap \ \Omega|}{|M \ \cup \ \Omega|}. \tag{8}$$

4. EXPERIMENTAL RESULTS

4.1 Classification

Table 2 summarizes the mean classification effectiveness obtained by the six different models. A total of four metrics, as outlined in Section 3.4.1, were used to evaluate these models. In terms of mean accuracy, all models obtained values between 85.6% and 90.8%. Likewise, the mean precision, recall, and F1 score values vary between 84.2% and 90.4%, 84.8% and 90.4%, and 84.2% and 90.0%, respectively. Regardless of the metric selected, all models come with an effectiveness of at least 84.2%. The latter observation implies that, in the case of classification, the predictions made are not biased, even though there is some class imbalance in the dataset used. For AlexNet and VGG-19, the standard version is more effective than the modified version, while VGG-16 shows the opposite behaviour. However, the differences in effectiveness between a model and its modified version are less than 2%. Furthermore, the modified versions have a slightly larger standard deviation. In conclusion, the best scoring model on all metrics is Modified VGG-16, with Modified AlexNet having the lowest effectiveness.

*https://scikit-learn.org/stable/

Table 2. Mean (standard deviation) of the classification effectiveness of the different models used.

Model	Accuracy	Precision	Recall	F1 score
AlexNet	0.860 (0.112)	0.846 (0.128)	0.846 (0.116)	0.844 (0.121)
VGG-16	0.900 (0.109)	0.890 (0.116)	0.884 (0.116)	0.886 (0.118)
VGG-19	0.896 (0.117)	0.890 (0.109)	0.894 (0.099)	0.888 (0.112)
Modified AlexNet	0.856 (0.127)	0.842 (0.131)	0.848 (0.129)	0.842 (0.131)
Modified VGG-16	**0.908 (0.123)**	**0.904 (0.116)**	**0.904 (0.110)**	**0.900 (0.120)**
Modified VGG-19	0.894 (0.117)	0.886 (0.119)	0.882 (0.116)	0.886 (0.119)

4.2 Segmentation

CAM can be used to understand how a deep learning model makes a prediction. To do so, it generates a heatmap that highlights the most informative regions associated with a class of interest. In order to quantitatively analyze how well our classification models recognize brain tumors and determine their type, we perform binarization of the obtained class activation maps by making use of different threshold values, hereby generating proxy segmentation masks that can be compared to the ground truth segmentation masks. The mIoU values obtained for each classification model, using different threshold values, can be found in Table 3.

Among the five threshold values used, 200/255 comes with the highest mIoU, followed by Q95 by a slight margin. Since Q95 is determined using the pixel values in an image instead of being user-defined, it can be considered a more general criterion. Moreover, Modified VGG-16 recorded the highest mIoU among the different trained networks. In addition, compared to the three standard models, we observe an increase in mIoU for the modified models.

Furthermore, as shown in Table 4, when considering the mIoU per class for a threshold value of 200/255, we observe an inequality. In particular, for all instances, meningioma has the highest mIoU, followed by glioma, and then by pituitary tumor. Moreover, as can be seen in Table 5, 49.3% of the pituitary tumors that are correctly classified by Modified AlexNet have an IoU less than 0.01. The same observation regarding the IoU can be made for 85.3% and 69.9% of the pituitary tumors tha are correctly classified by Modified VGG-16 and Modified VGG-19, respectively. In this respect, it is worth mentioning that an IoU of less than 0.01 points to a badly segmented tumor, given that the similarity between a CAM-based mask and the corresponding ground truth is then less than 1%. In addition, as can be seen in Figure 3, none of the pituitary tumors have an IoU higher than 23%.

For comparison purposes, the mIoU obtained by U-Net can be found in the bottom rows of Table 3 and Table 4. We observe that the mIoU obtained by our classification models is significantly lower than the mIoU obtained by U-Net. For example, the mIoU obtained by Modified VGG-16 and Modified VGG-19 at a threshold value of 200/255 is about six times lower than the mIoU obtained by U-Net. In this context, however, it is important to point out that a direct comparison needs to be done in a careful way. Indeed, from a loss point-of-view, there are significant differences, with U-Net working with precise pixel-level loss and with our classification models working with label-level loss.

The distribution of the IoU per tumor class can be found in Figure 3, visualized by making use of histograms. The histogram of U-Net is skewed to the right, showing a significant number of instances that come with a high IoU. On the other hand, the modified networks have a high peak in the first bin, corresponding to an IoU in the range of 0 to 0.05. We observe that the segmentation ability of the classification models, through the use of CAM and as measured by IoU, is weak compared to the inherent segmentation ability of U-Net.

Table 3. mIoU values (standard deviation) obtained by the different classification models for different threshold values. For U-Net, the encoder used is enclosed in parentheses.

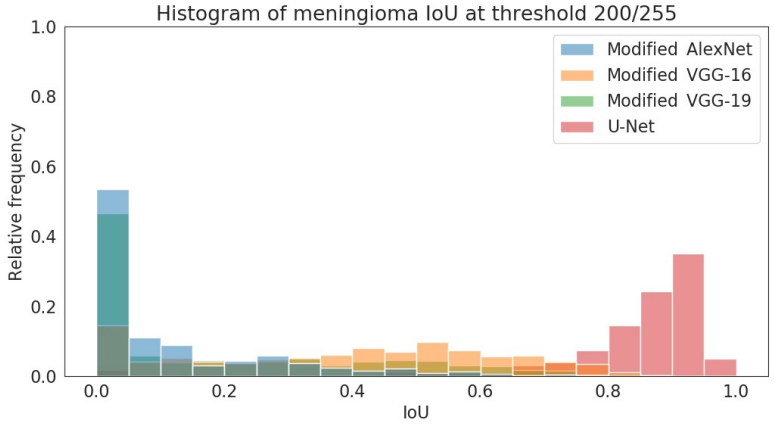
Networks	mIoU (for all classes)				
	200/255	225/255	Q90	Q95	Otsu
AlexNet	0.021 (0.008)	0.016 (0.007)	0.019 (0.009)	0.016 (0.007)	0.023 (0.008)
VGG-16	0.095 (0.027)	0.068 (0.022)	0.070 (0.015)	0.091 (0.027)	0.048 (0.008)
VGG-19	0.103 (0.028)	0.073 (0.020)	0.084 (0.017)	0.100 (0.028)	0.061 (0.012)
Modified AlexNet	0.064 (0.013)	0.059 (0.009)	0.057 (0.013)	0.059 (0.013)	0.038 (0.009)
Modified VGG-16	**0.122 (0.026)**	0.103 (0.019)	0.075 (0.009)	0.103 (0.016)	0.062 (0.014)
Modified VGG-19	0.108 (0.038)	0.091 (0.034)	0.084 (0.019)	0.101 (0.033)	0.058 (0.010)
U-Net (VGG-16)			0.666 (0.041)		
U-Net (VGG-19)			0.673 (0.038)		

Table 4. mIoU values (standard deviation) obtained by the different classification models per class for a fixed threshold value of 200/255. For U-Net, the encoder used is enclosed in parentheses.

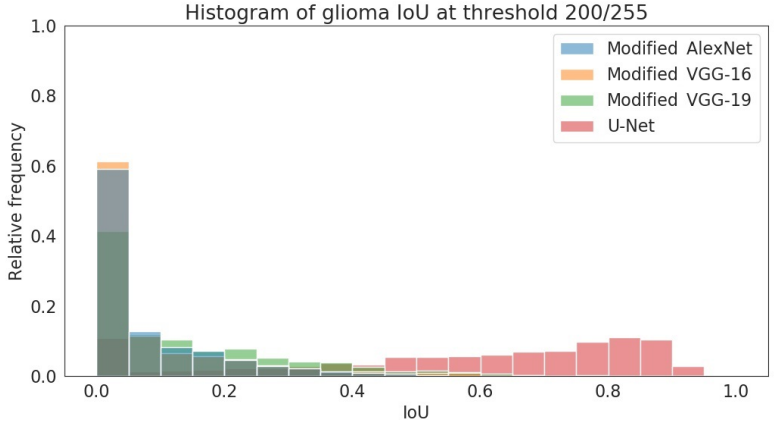
Networks	mIoU (threshold: 200/255)			
	All classes	Meningioma	Glioma	Pituitary
AlexNet	0.021 (0.008)	0.052 (0.027)	0.013 (0.004)	0.014 (0.011)
VGG-16	0.095 (0.027)	0.249 (0.096)	0.068 (0.028)	0.004 (0.005)
VGG-19	0.103 (0.028)	0.220 (0.077)	0.081 (0.021)	0.010 (0.005)
Modified AlexNet	0.064 (0.013)	0.107 (0.034)	0.061 (0.009)	**0.021 (0.013)**
Modified VGG-16	0.122 (0.026)	**0.369 (0.108)**	0.071 (0.010)	0.009 (0.008)
Modified VGG-19	0.108 (0.038)	0.183 (0.130)	**0.111 (0.021)**	0.013 (0.012)
U-Net (VGG-16)	0.666 (0.041)	0.824 (0.027)	0.545 (0.060)	0.732 (0.046)
U-Net (VGG-19)	0.673 (0.038)	0.835 (0.034)	0.559 (0.054)	0.724 (0.042)

Table 5. Proportion of images with an IoU ≤ 0.01, as obtained by the different classification models for different classes at a threshold value of 200/255.

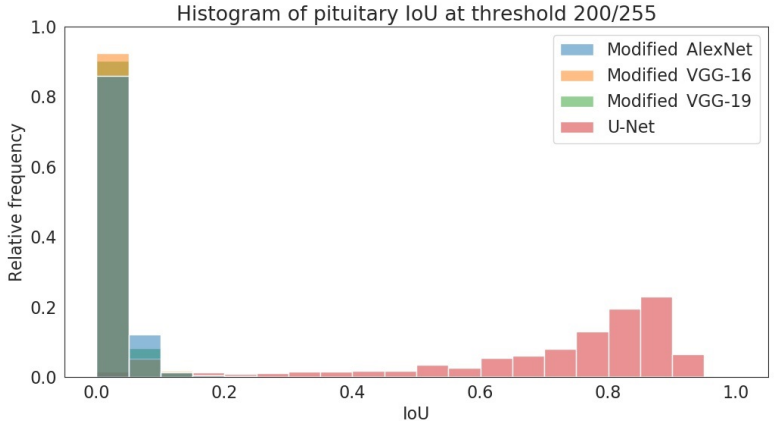
Networks	Proportion (threshold: 200/255)		
	Meningioma	Glioma	Pituitary
AlexNet	**0.604**	**0.791**	0.685
VGG-16	0.205	0.478	**0.916**
VGG-19	0.247	0.361	0.792
Modified AlexNet	0.446	0.452	0.493
Modified VGG-16	0.105	0.489	0.853
Modified VGG-19	0.368	0.316	0.699
U-Net (VGG-16)	0.007	0.097	0.022
U-Net (VGG-19)	0.017	0.095	0.011



(a) Meningioma

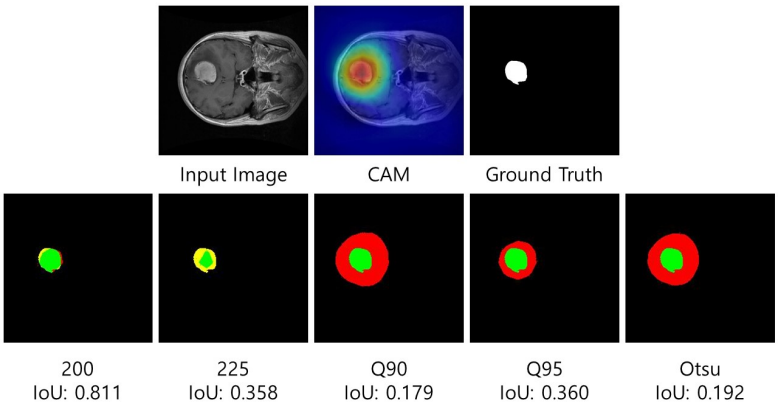


(b) Glioma

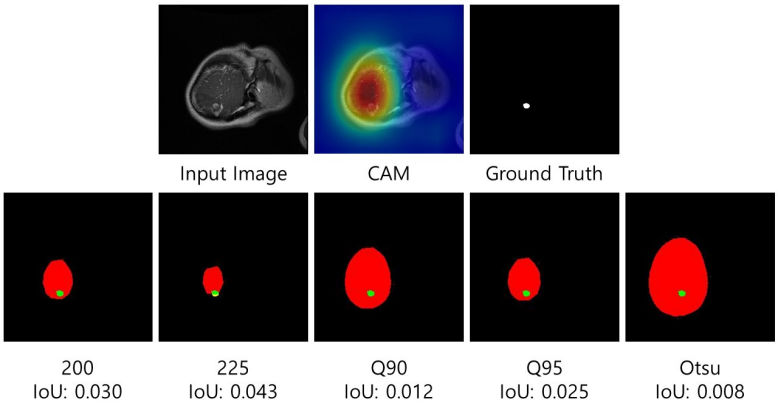


(c) Pituitary tumor

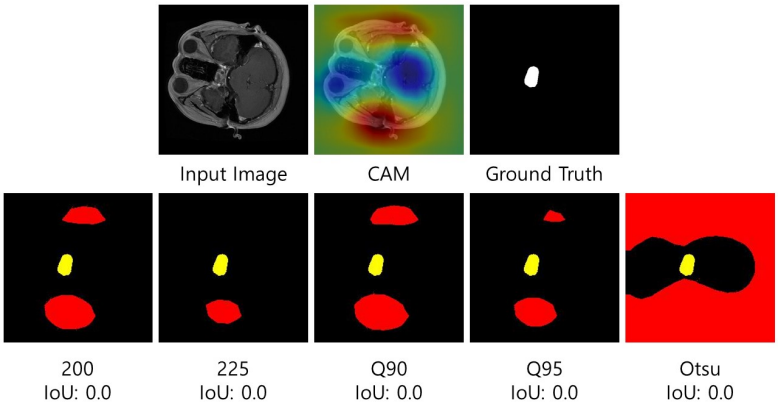
Figure 3. Histograms of the number of images binned according to their IoU score per tumor class, for both the modified models and U-Net (using VGG-19 as the encoder). Note that relative frequency, as used on the y-axis, refers to the image count.



(a) Example of CAM-based segmentation masks obtained for meningioma, showing corresponding threshold and IoU values.



(b) Example of CAM-based segmentation masks obtained for glioma, showing corresponding threshold and IoU values.



(c) Example of CAM-based segmentation masks obtained for pituitary tumor, showing corresponding threshold and IoU values.

Figure 4. Qualitative analysis of mask generation per threshold value (using Modified VGG-16). The colors in a generated mask mean the following: green denotes correctly detected tumor pixels, black denotes correctly detected background pixels, red denotes background pixels incorrectly detected as tumor pixels, and yellow denotes tumor pixels incorrectly detected as background pixels.

5. DISCUSSION

CAM is a visualization tool that makes it possible to output a heatmap, highlighting those areas within an input image that are the most informative for a predicted class (that is, the most informative regions that made the model predict a certain type of brain tumor). As such, assuming correct classification results, it is reasonable to expect that CAM is also able to highlight the location of a brain tumor, suggesting that a classification model can both classify brain tumors and determine their location, and where the last capability can be quantified by making use of IoU.

The six CNN-based classification models used in our study - AlexNet, VGG-16, VGG-19, Modified AlexNet, Modified VGG-16, and Modified VGG-19 - all come with a high accuracy between 85.6-90.8%. Furthermore, modifications to the standard classification models make it possible to increase the mIoU without significantly affecting the classification effectiveness. This indicates that the segmentation ability has been improved, and this with the help of one or more additional convolutional layers.

Given our experimental results, we observe a difference in mIoU per tumor class. For example, although meningioma makes up for the smallest portion of our dataset (23.1%), the best results are obtained for this type of tumor. We can identify two possible explanations for this observation. The first reason may be linked to the location of the different types of tumors. Meningioma are, for instance, located at the border of the brain. When analyzing badly segmented images, we could observe that the area surrounding the brain was highlighted, possibly pointing to models that were trained to look around the brain to find meningioma and differentiate these tumors from the background. On the other hand, glioma and pituitary tumors are located in the center of the brain. In the majority of cases, and this for both classes, we found that either the entire brain is highlighted or that other organs, such as eyes or ears, were highlighted. Indeed, as shown in Figure 4(c), both the left and the right ear are highlighted in red as pituitary tumor. As such, the models are having difficulties in distinguishing these kinds of tumor, resulting in a low mIoU. The other reason for a low mIoU is that the area segmented by CAM is large in comparison to the actual ground truth. Indeed, even if the model was able to correctly segment a tumor (through CAM), we could observe a significant difference between the generated segmentation mask and the ground truth mask, as illustrated in Figure 4(b). Therefore, not only when a tumor is misidentified, but also when it is less precisely recognized, this results in a low IoU.

In summary, CAM can provide insight into the decisions made by a predictive model. By applying a simple thresholding method to class activation maps, segmentation masks can be obtained in a straightforward way, which can then, in turn, be used to study the effectiveness of the model. Such an alternative method for quantitative model analysis can be helpful in the context of healthcare, where correct decision-making is crucial, given that decisions may be a matter of life or death. A model having a high mIoU and a high classification accuracy should be able to distinguish abnormalities at the right place and should be able to classify these abnormalities correctly, thus allowing for accurate diagnoses. If the classification effectiveness of a model is high but the mIoU is low, which we observe for our experiments in Section 4, careful attention is required when interpreting the results obtained, as the model does not seem to be looking at the correct location of the tumor when performing the classification.

6. CONCLUSIONS AND FUTURE RESEARCH

Using CAM as a proxy for image segmentation needs to be done in a cautious way, given that CAM learns from class labels and not from pixel-wise labels (as available in a binary segmentation mask). Nonetheless, by using CAM as a proxy for image segmentation, we found that (1) a predictive model does not always focus on the correct areas in an image to perform classification and that (2) a high classification effectiveness can still be paired with a low segmentation effectiveness (given the observed classification effectiveness of 90.8% versus an mIoU of only 12.2% in case of Modified VGG-16. Therefore, healthcare practitioners applying CNN-based classification models to medical images need to be careful when interpreting the obtained model results.

In the future, the following additional experiments can be performed:

Class imbalance – Medical datasets are often not balanced, possibly leading to biased predictions. To mitigate class imbalance, the loss function could be modified to give each class a different weight. Specifically, more weight could be given to the minority class and less weight to the majority class. This weighted loss method could thus help in treating unbalanced data in a more balanced way.[19]

Classification models – In this study, only two types of classification models were used, namely AlexNet and VGG. These types of classification models were released in 2012 and 2014, respectively. In future research, we plan to investigate the use of ResNet,[20] DenseNet,[21] and other recently released models.

Heatmap generation – As an alternative to CAM, methods such as Grad-CAM++ and Guided Attention Inference Networks could be applied. Grad-CAM++[22] produces heatmaps that help to segment complex objects (e.g., two ducks in an image) by fixing the weight values. On the other hand, Guided Attention Inference Networks[23] make it possible to obtain an improved segmentation effectiveness by learning not only classes but also a small part of the segmentation mask in the model. Moreover, future research efforts may consider the use of Local Interpretable Model-agnostic Explanations (LIME)[24] and SHapley Additive exPlanations (SHAP).[25]

Mask refinement –The size of the original heatmap corresponds to the size of the feature map before the GAP layer (14×14). Afterwards, this heatmap is up-sampled to the size of the input image (224×224) for visualization purposes. Due to the large resolution difference, the generated mask may have difficulties in distinguishing tumor from background in a fine-grained way. As an alternative, mask refinement through the use of a Conditional Random Field (CRF)[5] could be adopted. That way, a model can learn the boundary of the mask from a given image, thus helping in generating a more precise segmentation mask.

Acknowledgments

The research and development activities described in this paper were funded by Ghent University Global Campus (GUGC), Incheon, Korea.

REFERENCES

[1] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "ImageNet Classification with Deep Convolutional Neural Networks," in [*Advances in Neural Information Processing Systems 25*], Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., eds., Curran Associates, Inc. (2012).

[2] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al., "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning," *arXiv preprint arXiv:1711.05225* (2017).

[3] Bien, N., Rajpurkar, P., Ball, R. L., Irvin, J., Park, A., Jones, E., Bereket, M., Patel, B. N., Yeom, K. W., Shpanskaya, K., et al., "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet," *PLoS Medicine* **15**(11), e1002699 (2018).

[4] Han, S., Kim, M., Lim, W., Park, G., Park, I., and Chang, S., "Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm," *Journal of Investigative Dermatology* **138**(7), 1529–1538 (2018).

[5] Nguyen, H.-G., Pica, A., Hrbacek, J., Weber, D. C., La Rosa, F., Schalenbourg, A., Sznitman, R., and Cuadra, M. B., "A novel segmentation framework for uveal melanoma in magnetic resonance imaging based on class activation maps," in [*International Conference on Medical Imaging with Deep Learning*], 370–379 (2019).

[6] Kiani, A., Uyumazturk, B., Rajpurkar, P., Wang, A., Gao, R., Jones, E., Yu, Y., Langlotz, C. P., Ball, R. L., Montine, T. J., et al., "Impact of a deep learning assistant on the histopathologic classification of liver cancer," *npj - Digital Medicine* **3**(1), 1–8 (2020).

[7] Kim, M., Han, J., Hyun, S., Janssens, O., Van Hoecke, S., Kee, C., and De Neve, W., "Medinoid: Computer-Aided Diagnosis and Localization of Glaucoma Using Deep Learning," *Applied Sciences* **9**(15), 3064 (2019).

[8] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A., "Learning Deep Features for Discriminative Localization," in [*2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*], 2921–2929, IEEE (2016).

[9] Waring, J., Lindvall, C., and Umeton, R., "Automated machine learning: Review of the state-of-the-art and opportunities for healthcare," *Artificial Intelligence in Medicine* (2020).

[10] Anwar, S. M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., and Khan, M. K., "Medical Image Analysis using Convolutional Neural Networks: A Review," *Journal of Medical Systems* (2018).

[11] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," in [*Proceedings of IEEE International Conference on Computer Vision*], 618–626 (2017).

[12] Cheng, J., *Brain tumor dataset* (2017 (accessed July 28th, 2020)). https://doi.org/10.6084/m9.figshare.1512427.v5.

[13] Simonyan, K. and Zisserman, A., "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556* (2014).

[14] Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., and Liu, C., "A Survey on Deep Transfer Learning," in [*Artificial Neural Networks and Machine Learning – ICANN 2018*], 270–279, Springer International Publishing (2018).

[15] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L., "ImageNet: A large-scale hierarchical image database," in [*2009 IEEE conference on computer vision and pattern recognition*], 248–255, IEEE (2009).

[16] Ronneberger, O., Fischer, P., and Brox, T., "U-Net: Convolutional Networks for Biomedical Image Segmentation," in [*Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*], 234–241, Springer International Publishing (2015).

[17] Milletari, F., Navab, N., and Ahmadi, S.-A., "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation," in [*2016 Fourth International Conference on 3D Vision (3DV*], 565–571, IEEE (2016).

[18] Otsu, N., "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man, and Cybernetics* **9**(1), 62–66 (1979).

[19] Johnson, J. M. and Khoshgoftaar, T. M., "Survey on deep learning with class imbalance," *Journal of Big Data* **6**, 27 (Mar 2019).

[20] He, K., Zhang, X., Ren, S., and Sun, J., "Deep Residual Learning for Image Recognition ," in [*Proceedings of IEEE International Conference on Computer Vision*], 770–778 (2016).

[21] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q., "Densely Connected Convolutional Networks," in [*Proceedings of IEEE International Conference on Computer Vision*], 4700–4708 (2017).

[22] Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N., "Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks," in [*2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*], 839–847 (2018).

[23] Li, K., Wu, Z., Peng, K.-C., Ernst, J., and Fu, Y., "Tell Me Where to Look: Guided Attention Inference Network," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*], 9215–9223 (2018).

[24] Ribeiro, M. T., Singh, S., and Guestrin, C., ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in [*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*], *KDD '16*, 1135–1144, Association for Computing Machinery, New York, NY, USA (2016).

[25] Lundberg, S. M. and Lee, S.-I., "A Unified Approach to Interpreting Model Predictions," in [*Advances in Neural Information Processing Systems*], Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., eds., **30**, 4765–4774, Curran Associates, Inc. (2017).